Evaluación de resultados clínicos (II): Las medidas de la significación clínica o los tamaños del efecto

Ioseba Iraurgi
Deusto–Salud, I+D+i en Psicología Clínica y de la Salud.
Universidad de Deusto. Bilbao

Si tu experimento necesita estadística,
deberías haber hecho uno mejor.
Ernest Rutherford

Resumen: Como bien planteaba Glass *'la significación estadística es lo menos interesante de los resultados. Éstos se deberían describir en términos de la magnitud de la medida; no sólo cómo afecta el tratamiento a los sujetos, sino cuánto les afecta'*. El término tamaño del efecto (TE), un concepto elaborado por Jacob Cohen, se refiere a la magnitud de una medida de resultado o a la fuerza de una relación entre dos variables. En este artículo se propone una descripción de varios índices del tamaño del efecto agrupados en tres clases o familias: 1) la familia de los índices basados en diferencias (d de Cohen, g de Hedges, delta de Glass y la Diferencia de Riesgos; 2) la familia de las correlaciones; y 3) la familia de las razones (el riesgo relativo y la odds ratio).

Palabras clave: Tamaño del Efecto, Diferencia de Medias Estandarizada, Razón de Riesgos, Precisión del efecto, Significación clínica.

Summary: As good was proposing Glass, *'Statistical significance is the least interesting thing about the results. You should describe the results in terms of measure of magnitude - nor just, does a treatment affect people, but how much does it affect them'*. The Effect Size (ES), a concept developed by Jacob Cohen, is generally used to refer to the magnitude of an outcome result or to the strength of the relationship between two variables. In this paper we provide a sampling of ES indices in three general classes: 1) the difference family (Cohen´s d, Hedges´s g, Glass´s Delta and the Risk Difference); 2) the correlation family; and 3) the ratio family (Relative Risk and Odds Ratio).

Keywords: Effect Size, Standardized Mean Difference, Risk Ratio, Effect Precision, Clinic significance.

Siguiendo con el estilo iniciado en la primera entrega de esta serie de artículos metodológicos sobre la evaluación de resultados clínicos, hemos iniciado este artículo con una cita; en este caso del físico–químico británico Ernest Rutherford (1881–1937), considerado el padre de la física nuclear. El propósito no es minusvalorar la importancia de la estadística, sino redireccionar su sentido y papel dentro del proceso de investigación. La estadística es una herramienta fundamental, siempre y cuando sepamos qué efecto estamos buscando y hayamos seleccionado los elementos de diseño y planificación del experimento necesarios para

establecer las líneas de causalidad que queremos probar. De otro modo, cabría esperar lo que ya Fisher propuso en otra de sus célebres reflexiones: *'Pedir ayuda a un estadístico una vez que el experimento ya fue realizado puede no ser más que pedirle efectuar una autopsia: posiblemente, lo único que pueda hacer sea decir qué el experimento murió'*.

En la investigación clínica, en el caso concreto de la valoración de la bondad terapéutica de dos modalidades de tratamiento, el propósito fundamental es probar que una de estas modalidades es superior a la otra (en eficacia terapéutica, seguridad, comodidad de uso, reducción de efectos adversos, etc.). Por lo general, comparamos una nueva modalidad con otra ya probada, o utilizada de forma indicada en la práctica habitual, que se toma como tratamiento de comparación o control. Nuestra hipótesis es que la nueva modalidad será mejor que la ya conocida y utilizada de forma estándar; y lo que esperamos conocer es cuánto mejor es ésta, ya que de ello surgirá una nueva indicación o elección terapéutica. De este modo, sabemos qué estamos buscando y, además, en la formulación de nuestro estudio tenemos que anticipar cuánto mejor esperamos que sea la nueva modalidad de tratamiento. La estimación de este cuánto estará más relacionado con el logro de una diferencia clínicamente significativa, que con la obtención de una diferencia estadísticamente significativa. Recordemos lo ya revisado en la entrega anterior (Iraurgi, 2009), un resultado estadísticamente significativo sólo indica que es poco probable que la relación encontrada entre las variables sea debida al azar y, por tanto, esa relación puede ser aceptada como real. No obstante, la significación estadística no proporciona información sobre la fuerza de la relación (tamaño del efecto) o si la relación es reveladora (significación clínica). Por tanto, lo que nos interesa de forma prioritaria es elegir un indicador que nos aproxime a la significación clínica a través de la estimación de la magnitud o tamaño del efecto.

Se entiende por 'efecto' la consecuencia que produce una intervención o exposición sobre el evento de interés; también llamado variable de resultado, logro final, o en su acepción anglosajona, 'outcome'. El efecto puede ser beneficioso (obtención de una diferencia a favor del tratamiento que ponemos a prueba o experimental), perjudicial (las diferencias detectadas son a favor del tratamiento de control o, visto de otro modo, en contra del tratamiento experimental), y también podría ser nulo (no hay diferencias entre ambas modalidades terapéuticas). Esta primera observación es lo que se conoce como 'dirección del efecto' que, como hemos dicho, suele venir ya especificada en la hipótesis de estudio. Pero lo que nos interesa con muestro ensayo, y precisamente para eso lo realizamos, es estimar un valor numérico que de cuenta de la magnitud del efecto, un indicador que permita cuantificar la fuerza de la asociación entre la intervención y la respuesta lograda. Cuanto mayor sea este efecto, cuanto más fuerza exprese la asociación entre variables, cuanto mayor sean las diferencias de resultado obtenidas entre ambas modalidades de tratamiento, entonces, existirá una mayor probabilidad de que ese efecto sea debido a la intervención y, en ese caso, la probabilidad de que dicho efecto sea debido al azar tenderá a ser pequeña; es decir, el valor 'p' de significación estadística asociado será inferior a los límites convencionalmente aceptados ($p \leq 0{,}05$ o inferiores). En definitiva, la significación estadística nos permitirá concluir que el efecto alcanzado no se debe al azar, sino que es producto de la acción de la intervención terapéutica ofrecida; pero realmente será la magnitud del efecto logrado el que nos proporcionará la información necesaria para decidir si los resultados obtenidos alcanzan o no la significación clínica.

En la metodología estadística no existe un indicador único, universal, para poner a prueba la significación clínica, sino varios. Las contribuciones de diversos investigadores han sido prolijas y se ha sugerido clasificarlas en tres tipos o familias de medidas del tamaño del efecto atendien-

do al tipo de relacione entre variables observado. A este respecto, existe un conjunto de indicadores que valoran la fuerza de la asociación entre variables y a él pertenecería toda la gama de coeficientes de asociación y/o correlación (*r* de Pearson, r_S de Sperman, coeficiente de determinación, coeficientes tetracóricos, etc.); todos ellos son agrupados bajo la denominación de 'familia *r* del tamaño del efecto'. Otros indicadores estiman la magnitud de la diferencia entre las medias de los efectos obtenidos bajo un tratamiento en comparación con un grupo control, donde la variable de resultado es una medida de tipo continuo; estos indicadores pertenecen a la 'familia d del tamaño del efecto', y son característicos de la misma la conocida d de Cohen, la $\triangle$ (delta) de Glass y la g de Hedge. Un tercer tipo de familia son las derivadas de la estimación del riesgo que se obtienen a través de tablas 2x2, al que pertenecen la Odds–Ratio (OR), el Riesgo Relativo (RR), la Reducción Relativa del Riesgo (RRR), la Diferencia de Riesgos (DR) o Riesgo Atribuible, el Número Necesario de pacientes a Tratar (NNT) o el Número Necesario de pacientes a Dañar (NND). Todos estos índices nos ofrecen un indicador del tamaño del efecto observado entre las variables sometidas a examen, pero se hayan afectados de un margen de error, de una determinada precisión de medida, que en definitiva será expresión de su significación estadística. El cálculo de los intervalos de confianza de estos índices nos permitirán tomar decisiones sobre su significación estadística y clínica. En este artículo, se pretende ofrecer una introducción a los índices propuestos, y ofrecer orientaciones que faciliten la interpretación y la comprensión de la significación clínica.

Tamaño del Efecto (TE) y familias de clasificación

El tamaño del efecto (effect size), un concepto elaborado por Jacob Cohen (1962, 1969–88) en su ya clásico texto sobre el análisis del poder estadístico, es generalmente utilizado en la investigación biomédica y del comportamiento para referirse a la magnitud de una medida de resultado o a la fuerza de una relación entre dos variables, por lo general una independiente y la otra dependiente (Coe, 2002; Carson, 2004).

Una de las ventajas más importantes de los índices del TE —como podremos apreciar un poco más adelante— es que son transformaciones a una escala común, a una escala carente de unidades (las correlaciones, las diferencias estandarizadas, las estimaciones de riesgos, …), de modo que los resultados de diferentes estudios pueden ser directamente comparables. Esta característica es la que les hace imprescindibles al realizar estudios meta–analíticos (). De hecho, el desarrollo de estos estadísticos ha de ser atribuido sobre todo a los teóricos y expertos de esta técnica (Cohen, Glass, Rosenthal,…), muy útil a la hora de resumir, estructurar y acumular el conocimiento científico obtenido en investigaciones empíricas que utilizan metodología estadística (Rosenthal, 1991; Cooper y Hedge, 1994; Sánchez y Ato, 1989; Sánchez–Meca, Marín–Martínez y Chacón–Moscoso, 2003; González–Ramírez y Botella, 2006).

La elección de las diferentes pruebas estadísticas disponibles está supeditada a las características de las variables que estemos analizando (Iraurgi, 2000) y, aunque para cada una de estas pruebas existe un TE (e incluso varios modos de calcular el TE para una misma prueba), es posible la transformación entre los diferentes tipos. Así, es común en los estudios meta–analíticos que todos los TE diferentes (*d* de Cohen, *f, h,* $\triangle$*, q, W*, etc) se conviertan, por ejemplo, a coeficientes de correlación *r* de Pearson (Rosenthal, 1991; Marín–Martínez y Sánchez–Meca, 1996; Sánchez–Meca y Marín–Martínez, 2001; González–Ramírez y Botella, 2006).

Familia *r* del tamaño del efecto: el caso de la correlación entre variables

Uno de los coeficientes más conocidos en estadística es la *r* de Pearson, también llamado coeficiente de correlación producto–momen-

to. Se calcula cuando tratamos de establecer el grado de asociación entre dos variables medidas en escala continua o de razón; para variables de intervalo se utiliza el coeficiente de correlación por rangos de Spearman (r_s), existiendo un gran número de coeficientes (el Biserial–puntual, la Q de Yule, etc.) que variaran en función de la combinación de la escala de las variables utilizadas.

En todos los coeficientes de asociación referidos los valores asumibles oscilan entre –1 y +1 siendo el valor cero expresión de la ausencia de asociación, y aumentando el grado o magnitud de la asociación entre las variables a medida que se aproximan al valor unidad; el signo positivo o negativo informará sobre el sentido de la asociación. De este modo, un coeficiente de correlación igual a 1 (ó –1) indica que para cada valor de la variable X le corresponde un valor de la variable Y, existiendo, por tanto, una asociación perfecta.

En la Tabla 1 se presentan las directrices generales para la interpretación de la magnitud del efecto de cinco medidas que se presentarán en este artículo de revisión. Se presentan una serie de valores que expresarían un punto de inflexión a partir del cual se valoraría el TE a través de una clasificación valorativa: efecto pequeño, mediano, grande. Cohen (1988) propuso esta clasificación cualitativa a partir de la revisión de los TE observados comúnmente en la investigación en general, es decir, a partir de los valores más típicos. Por esta razón, algunos autores recomiendan poner en relación los valores propuestos por Cohen en referencia a esta tipicidad. De este modo, los valores de correlación en trono a 0,30 serían los más típicos en investigación; valores por debajo de ese valor serían valores más pequeños que el típico, y valores por encima de 0,50 serían valores más grandes que lo típico. Es decir, la decisión sobre la magnitud de un coeficiente de correlación se establece a partir de lo que comúnmente se observa en la investigación del área. En algunos casos, por ejemplo enfermedades de muy baja prevalencia, puede encontrarse un coeficiente con un valor de 0,20 y ser considerado muy alto, ya que lo típico en la investigación de esa enfermedad se sitúa en valores de 0,08. A este respecto, Lipsey (1990) propone que para determinar un tamaño del efecto en la planificación de una investigación es muy útil considerar la distribución de los tamaños del efecto de las investigaciones anteriores similares. Sugiere un procedimiento para sustituir la convención de tamaños del efecto propuesta por Cohen (1988). Si existe un meta–análisis sobre el tema puede utilizarse como tamaño del efecto pequeño la media o mediana de la distribución de puntuaciones de tamaños del efecto que quedan por debajo del percentil 33; como tamaño del efecto medio la media o mediana del 34% central de las puntuaciones de tamaños del efecto; y como tamaño del efecto alto la media o mediana de las magnitudes del efecto que estén por encima del percentil 67.

Otra Forma de estimar el grado de asociación o magnitud del efecto es a partir del coeficiente de determinación R^2. Este se calcula elevando al cuadrado el coeficiente de correlación y expresa el porcentaje de varianza conjunta entre las variables consideradas. De este modo, una correlación de 0,50, considerada como más grande que la típica, expresaría un tan solo una varianza común del 25%, o en una relación de causalidad, la variable independiente solo llegaría a predecir un 25% de la variable respuesta. Ahora bien, debemos tener en cuenta que en las ciencias del comportamiento y de la salud las explicaciones posibles son multicausales, de modo que un determinado fenómeno comparte su varianza con gran número de variables. Por todo ello, acogernos a la propuesta de Cohen o, quizá más acertadamente a la de Lipsey, se hace necesario para ordenar y establecer nuestras decisiones.

Familia *d* del tamaño del efecto: el caso de la diferencia de medias

Mediante esta familia de estimadores se pretende estimar el grado de generalidad

Tabla 1

Interpretación de la fuerza (tamaño) del efecto de los indicadores más usuales.

Interpretación general de la fuerza del efecto	Familia de las correlaciones *r*	Familia de las diferencias *d*	Familia de Estimadores del Riesgo			
			OR	RR	DR o RAR	NNT
Perfecta	1,00	≥ 3,5 (∝)	∝	∝	1,00	1,0
Casi perfecta	0,90	2,50	360,00	19,00	0,90	1,1
Mucho mayor que el típico	0,70	1,20	32,00	5,70	0,70	1,4
Más grande que el típico	0,50	0,80	9,00	3,00	0,50	2,0
Medio o típico	0,30	0,50	3,50	1,90	0,30	3,3
Más pequeño que el típico	0,10	0,20	1,50	1,20	0,10	10
Ausencia de efecto	0,00	0,00	1,00	1,00	0,00	∝

r: Correlación de Pearson; **d**: de Cohen; **OR**: Odds Ratio; **RR**: Riesgo Relativo; **DR**: Diferencia de Riesgos, Riesgo Atribuible o Reducción Absoluta de Riesgo (**RAR**); **NNT**: Número Necesario de pacientes a Tratar

poblacional de un efecto a partir de la diferencia que se observa entre dos medias, bien de un mismo grupo en dos momentos temporales (lo que se conoce como comparación intra–sujetos), bien de dos grupos diferentes en un mismo momento temporal (comparación inter–sujetos), o bien de dos grupos diferentes en dos momentos temporales diferentes (comparación mixta o inter–intra–sujetos). En estos casos la variable independiente está medida en una escala nominal (de tipo binario o dicotómica, admite dos niveles de clasificación: experimental vs control) y la variable dependiente o de respuesta es de tipo continua (escala de razón). El cálculo del TE consiste en restar las medias de ambos grupos y dividir este resultado por una medida de variabilidad, una Desviación Estándar (DE). Los diferentes índices de esta familia del TE se diferencian, precisamente, en qué formula de la variabilidad utilicen (Tabla 2), pero todas ellas coinciden en ser una diferencia de medias estandarizada. Por tanto, el TE de la diferencia de medias estandarizadas varia de menos a más infinito, siendo el valor 0 indicativo de ausencia de efecto (la media de ambos grupos es la misma), de modo que a medida que las diferencias crecen la magnitud del efecto se hace cada vez más grande, si bien en las ciencias del comportamiento rara vez el valor del índice del TE es superior a 1.

Existen diferentes índices del TE incluidas en la familia d (citas); si bien las más características son la △ (Delta) de Glass, la *g* de Hedges y la *d* de Cohen (Tabla 2). En las tres fórmulas el numerador es la diferencia de la media del grupo de tratamiento (M_{Tto}) menos la media del grupo de control (M_{Ctrol}); si la diferencia es positiva indica que el tratamiento supera al control, y si es negativa que la puntuación del tratamiento es inferior al control. Para la Delta de Glass el denominador es la DE del grupo control; para la d de Cohen es la DE conjunta de las varianzas de ambos grupos, que cuando son muy similares equivale a una estimación media de la raíz cuadrada de las varianzas; y

TABLA 2
Índices del tamaño del efecto de la familia de las diferencias de medias y su equivalencia en 'r'

Índices	Algoritmo de Cálculo del Índice	Algoritmo de Calculo de la Desviación Estándar (denominador)	Tamaño del efecto en términos de correlación (Transformación)
'Δ' de Glass	$\frac{M_{Tto} - M_{Ctrol}}{S_{(Ctrol)}}$	$S_{(Ctrol)} = \sqrt{\frac{\sum (X_{Ctrol} - M_{Ctrol})}{n_{Ctrol} - 1}}$	$r = \frac{d}{\sqrt{d^2 + (1/pq)}}$
'd' de Cohen	$\frac{M_{Tto} - M_{Ctrol}}{Sa_{(comb)}}$	$Sa_{(Comb)} = \sqrt{\frac{S^2_{Tto} + S^2_{Ctrol}}{2}}$	Cuando los grupos experimental y control están formados por el mismo número de sujetos, entonces
'g' de Hedges	$\frac{M_{Tto} - M_{Ctrol}}{Sb_{(comb)}}$	$Sb_{(Comb)} = \sqrt{\frac{(n_{Tto} - 1) \cdot S^2_{Tto} + (n_{Ctrol} - 1) \cdot S^2_{Ctrol}}{n_{Tto} + n_{Ctrol} - 2}}$	$r = \frac{d}{\sqrt{d^2 + 4}}$ $d = \frac{2r}{\sqrt{1 - r^2}}$

M_{Tto} = Media del grupo de Tratamiento o Media post-tratamiento;
M_{Ctrol} = Media del grupo de Control o Media pre-tratamiento;
$S_{(Ctrol)}$ = Desviación Estándar del grupo Control;
Sa_{Comb}= Desviación Estándar conjunta de ambos grupos o medidas temporales;
S_{Tto} = Desviación Estándar del grupo de Tratamiento o post-tratamiento;
S_{Ctrol} = Desviación Estándar del grupo de Control o pre-tratamiento;
Sb_{Comb}= Media Cuadrática de Error intra-sujetos

'p'= n_{Tto}/N; y 'q'= n_{Ctrol}/N Corresponden a las proporciones de sujetos que pertenecen a los grupos experimental (n_{Tto}) y control (n_{Ctrol}), respectivamente;
N= Número total de casos

para la g de Hedges el denominador es la media cuadrática del error que se obtiene del cálculo de la prueba F de análisis de varianza.

Pongamos un ejemplo. Se ha realizado un estudio con 80 sujetos que se han distribuido al azar en dos grupos de 40. Al grupo experimental o de tratamiento se le ha ofrecido una nueva modalidad terapéutica obteniendo una M_{Tto}= 1,004 y una DE_{Tto}= 0,628; el grupo control, por su parte ha obtenido una M_{Ctrol}= 0,589 y una DE_{Ctrol}= 0,645; la media cuadrática de error ha resultado ser de 0,41, y la prueba F= 8,49 con una probabilidad asociada de $p \leq 0{,}005$. Este resultado se interpreta del modo siguiente: existe una probabilidad de equivocarnos en cinco veces de cada mil si admitimos que existen diferencias entre las puntuaciones alcanzadas por el grupo de tratamiento respecto al control; o dicho de otro modo, podemos aceptar que el grupo de tratamiento tiene un resultado más exitoso que el control siendo esta diferencia (de 0,415) estadísticamente significativas. Pero, ¿es realmente relevante esta diferencia?. Calculemos el TE:

a) △= (1,004–0,589) /0,645= 0,66;
b) d= (1,004–0,589) /√[(0,628²+0,645²)/2]= 0,65;
c) g= (1,004–0,589) /√0,41= 0,65.

Prácticamente los tres índices ofrecen el mismo valor del TE, de 0,65. ¿Y esto que significa?. Recordemos unas nociones fundamentales de estadística. Las puntuaciones Z estandarizadas consisten en restar a las puntuaciones de los sujetos (X_i) la media de su grupo de pertenencia (M) y dividir su resultado por la desviación estándar (DE) del grupo, de modo que obtendremos una nueva escala de medida con valores entre menos infinito y más infinito, con media cero (M_Z= 0) y desviación estándar de uno (DE_Z= 1). De este modo, dos variables medidas con distintos rangos de valores y/o unidades de medida podrían ser comparados, ya que las puntuaciones Z carecen de unidades de medida. Por otra parte, entre −1,96 y +1,96 puntuaciones Z se halla el 95% de las puntuaciones observadas en la muestra; a propósito, ¿les suena este valor?, es el que se utili-

za para estimar los intervalos de confianza del 95% [media$\pm z_{\alpha/2}\times EE$ → $M\pm 1{,}96\times(DE/\sqrt{n})$]. Entre ±1DE se hallan los valores que presentan el 68,26% de los sujetos de la muestra, y por encima de 1DE se localizan los valores del 15,86% de los sujetos de la muestra que presentan las puntuaciones más extremas; o dicho de otro modo, por debajo de 1DE se sitúan las puntuaciones obtenidas por el 84,14% de la muestra. Ahora volvamos sobre la diferencia de medias estandarizada. Al restar la media obtenida por el grupo experimental (GE) y el grupo control (GC), obtendremos la ganancia alcanzada por el GE frente al GC, y al ser dividida por la DE obtendremos una puntuación estandarizada con una lectura equivalente a las puntuaciones Z, de modo que si el valor del TE obtenido es de 0,80, por ejemplo, el porcentaje de personas del GC que se situaría por debajo de la media del GE sería de un 79% , es decir, podemos hacer una aproximación a los valores percentiles (columna 2 de la Tabla 3). En el caso de que el TE fuera 0,0 expresaría una falta de efecto (las medias de ambos grupos son iguales), de modo que el porcentaje de casos del GC que se situaría por debajo del GE sería del 50% —exactamente la misma proporción de casos del GE que se situaría por debajo del GC; ambas distribuciones se superponen—.

Tabla 3

Interpretación de la fuerza (tamaño) del efecto de las diferencias de medias (d de Cohen) y su equivalencia en valores de correlación (r).

	Tamaño del Efecto	Porcentaje del grupo control que se situaría por debajo de la media del grupo de tratamiento	Probabilidad de que una persona de grupo experimental será superior a una persona de control, si los dos son elegidos al azar	Equivalente en correlación 'r'	R^2 - Coeficiente de Determinación (% de varianza explicada)
Bajo	0,0	50,0%	0,0%	0,000	0,000
	0,1	54,0%	7,7%	0,050	0,002
	0,2	58,0%	14,7%	0,100	0,010
Medio	0,3	62,0%	21,3%	0,148	0,022
	0,4	66,0%	27,4%	0,196	0,038
	0,5	69,0%	33,0%	0,243	0,059
	0,6	73,0%	38,2%	0,287	0,083
	0,7	76,0%	43,0%	0,330	0,109
←Alto	*0,8*	79,0%	47,4%	0,371	0,138
	0,9	82,0%	51,6%	0,410	0,168
	1,0	84,0%	55,4%	0,447	0,200
	1,1	86,0%	58,9%	0,482	0,232
	1,2	88,0%	62,2%	0,514	0,265
	1,3	90,0%	65,3%	0,545	0,297
	1,4	91,9%	68,1%	0,573	0,329
	1,5	93,3%	70,7%	0,600	0,360
	1,6	94,5%	73,1%	0,625	0,390
	1,8	96,4%	77,4%	0,669	0,448
	2,0	97,7%	81,1%	0,707	0,500
	2,5	99,0%	87,7%	0,781	0,609
	3,0	99,9%	92,6%	0,833	0,693

Otro modo de estimación del TE para el caso de la diferencia entre dos medias provenientes de muestras independientes es el caso del índice *CLES* (Common Language Effect Size) propuesto por McGraw y Wong (1992) como una vía más simple de interpretación, ya que se expresa la magnitud de la diferencia en términos de un valor de probabilidad. En concreto, estima la probabilidad de obtener un valor de diferencias entre medias mayor que cero en una distribución normal cuya media es la diferencia observada entre ambas medias. Para su cálculo, y partiendo del ejemplo propuesto más arriba, se aplicaría la fórmula siguiente:

$$Z = \frac{|M_{Tto} - M_{Ctrol}|}{\sqrt{S^2_{tto} + S^2_{Ctrol}}} = \frac{|1{,}004 - 0{,}589|}{\sqrt{0{,}628^2 + 0{,}645^2}} = \frac{0{,}415}{1{,}044} = 0{,}460$$

obteniéndose un valor de Z. Seguidamente, se busca en las tablas de distribución normal tipificada la probabilidad de un valor menor al obtenido, de modo que p(z<0,460)= 0,6672; que se interpretaría del modo siguiente: la probabilidad de obtener una diferencia mayor que cero entre las puntuaciones de ambos grupos es de 0,6672; o, expresado de otro modo, en el 66,72% de las veces un sujeto extraído al azar del GE obtendrá un valor mayor que un sujeto extraído al azar del GC.

En la Tabla 3 se presentan las equivalencias para la interpretación de los distintos valores del TE d de Cohen, en valores percentiles (2ª columna), en valores de probabilidad (3ª columna) y se muestran, asimismo, sus equivalencias en valores r (4ª y 5ª columna), pudiéndose observar las formulas para la transformación de un índice en otro en la Tabla 2. Siguiendo con el ejemplo propuesto, para una *d*=0,66, le correspondería un valor *r*=0,31 [r= d/$\sqrt{(d^2+4)}$ = 0,66/$\sqrt{(0{,}66^2+4)}$].

Respecto a la interpretación del TE y su relevancia clínica, y en tanto que el texto que ahora afrontamos trata de aproximarnos al área de la evaluación de resultados clínicos, plantearemos como guía las observaciones realizadas por Kazdin y Bass (1989) en sus trabajos de revisión de la investigación en psicoterapia. Encontraron que en el caso de la investigación que pone a prueba un tratamiento frente a un placebo o condición de no tratamiento, sería necesario alcanzar una *d* ≥0,80 para considerar el TE como grande o clínicamente relevante, y una *d* ≥0,50 en el caso de investigación que pone a prueba un tratamiento experimental frente a otro con efecto activo, es decir, frente a otra modalidad terapéutica que ha mostrado efecto terapéutico.

TABLA 4
Presentación binomial del Tamaño del Efecto

	Éxito	Fracaso
Tratamiento	65,5	34,5
Control	34,5	65,5
Porcentaje de eficacia	+31%	-31%

Presentación binomial del tamaño del efecto

Existen otros modos de estimación del TE para el caso de la diferencia entre dos medias, como es el caso del valor contrapunto (Rosenthan y Rubin, 1994; citado en Valera y Sánchez, 1997), pero en aras a introducir la familia de índices del efecto siguiente nos parece oportuno citar la propuesta de Rosnow y Rosenthal (1996) sobre la 'presentación binomial del tamaño del efecto' (*BEST*, en su acepción en inglés: *Binomial Effect Size Display*). Estos autores han elaborado un procedimiento que puede ayudar a tomar una decisión al investigador sobre la utilidad práctica de un tratamiento. Esta forma de considerar la magnitud de un efecto consiste en transponer la correlación entre las variables en juego a una tabla 2×2 que incluya las proporciones de éxito (y de fracaso) del tratamiento. En el ejemplo que venimos utilizando, es posible convertir el TE *d* a un coeficiente *r* de Pearson, y a partir de este último índice confeccionar una tabla de proporciones de éxito y fracaso. Para ello, la conversión de r a *BESD* consiste en calcular la razón de éxito

en la condición experimental o de tratamiento de acuerdo con el algoritmo [0,50+(r/2)], y la razón de fracaso del tratamiento según [0,50–(r/2)]; en ambos casos se multiplica por cien para obtener porcentajes. Los porcentajes obtenidos se llevarían a una tabla 2×2 donde se clasificaría la respuesta obtenida (éxitos vs fracasos) respecto a la pertenencia a los grupos (tratamiento vs control). Así, volviendo al ejemplo propuesto, puede convertirse el tamaño del efecto a correlación de Pearson como hemos visto en párrafos anteriores: *d*= 0,66 equivale a una *r*= 0,31. Seguidamente calculamos el valor de BESD; [0,50+(0,31/2)= 0,655; 0,655×100= 65,5%] y, por tanto, la razón de fracaso del tratamiento será del 34,5% [(100–65,5); o bien: 0,50–(0,31/2) ×100= 34,5]. Llevados estos datos a una tabla de 2×2 (Tabla 4), se obtiene una información apreciable para la consideración de la relevancia clínica o utilidad práctica del tratamiento: el grupo tratado consigue un 31% más de eficacia que el grupo control, o bien, la intervención realizada en el grupo de tratamiento reduce el fracaso en un 31% respecto al grupo control.

Familia de las *Ratio* del tamaño del efecto: el caso de la estimación de riesgos

En el área de investigación biomédica existe todo un desarrollo de técnicas epidemiológicas que se han puesto al servicio de la evaluación terapéutica. Los diseños clásicos (cohortes, diseños de intervención, ensayos clínicos) se basan en la comparación de dos grupos diferenciados en función de la exposición o no a un factor 'causal' en los cuales se valora el desenlace de una determinada respuesta (salud–enfermedad, éxito–fracaso terapéutico, …) asociada al factor de exposición. La forma usual de trabajar con estas condiciones es a través de tablas 2x2 en las cuales, de forma estándar, se sitúan en las columnas la respuesta o variable dependiente, y en las filas los grupos de comparación o variable independiente. Uno de estos grupos representará la condición experimental o de tratamiento ya que es el que recibe la intervención que se espera logre un efecto o respuesta. Por su parte, la respuesta también se expresa de forma dicotómica, siendo una opción la respuesta positiva o buscada (el éxito o mejora terapéutica, por ejemplo) y la otra opción sería indicadora de una respuesta negativa o 'no positiva' (empeoramiento y/o no mejora). La definición de la respuesta positiva y negativa, del éxito o fracaso terapéutico, es una tarea que hay que hacer a priori, con la debida planificación de las condiciones de clasificación y la necesaria justificación teórica y clínica, para lo cual en la mayoría de las ocasiones nos basaremos en la evidencia acumulada. En la Figura 1 se presenta el esquema básico de este tipo de estructuración de tablas 2×2 y los algoritmos necesarios para calcular los índices de riesgo asociados a este tipo de estudios.

Para entender el sustrato a estos índices de riesgo, vamos a permitirnos introducir algunas nociones básicas de epidemiología, que pueden consultarse de forma más extensa en cualquier manual al uso (Armitage y Berry, 1997; Argimon y Jiménez, 2004; Irala, Martínez–González y Seguí–Gómez, 2008), pero que nos permitirán un desarrollo posterior más comprensible.

Una proporción es un cociente cuyo numerador está contenido en el denominador: en una muestra de 150 casos hay 98 hombres, por tanto la muestra tienen una proporción de hombres de 0,653 (98/150); si la proporción se multiplica por cien obtenemos el porcentaje (65,3%). Las proporciones carecen de unidades y su rango oscila entre 0 (0/n) y 1 (n/n). Índices epidemiológicos de este tipo son la prevalencia y la incidencia. En el caso de la prevalencia el denominador es el total de la población o muestra, y el numerador es el número de casos entre la población que presenta una determinada característica o efecto. Por su parte, la incidencia es la proporción de casos o acontecimientos nuevos (por ejemplo, casos que mejoran con un tratamiento), que aparecen en una población o muestra a lo largo de un periodo de tiempo. De este modo, la incidencia mide

la proporción de casos nuevos y, por tanto, requiere que exista un periodo de seguimiento de un grupo de personas en las cuales previamente no presentaban el efecto que se busca medir (en nuestro caso, el éxito terapéutico). En la epidemiología clásica las medidas de incidencia suelen estar referidas a la probabilidad de que ocurra un determinado efecto adverso o enfermedad, de ahí que también se denominen 'riesgos'.

Figura 1.- Algoritmos de cálculo de las medidas del efecto en tablas 2x2

		Respuesta (VD)			Ensayo de Intervención Cohortes
		Respuesta + Éxito Terap. (R+)	Respuesta – Fracaso (R-)	Total	Incidencia
Grupo (VI)	Tratamiento Intervención (T+) Expuesto	a	b	n_1	$\frac{a}{n_1}$
	Comparación Control (T-) No expuesto	c	d	n_2	$\frac{c}{n_2}$
	Total	m_1	m_2	N	$\frac{m_1}{N}$
Casos y Controles	Proporciones expuestas *(Odds)*	$\frac{a}{m_1}$	$\frac{b}{m_2}$		← Tipo de Estudios

Estudios de Cohortes y/o Ensayos de Intervención

$$I_1 = \frac{a}{n_1}; \quad I_2 = \frac{c}{n_2}$$

I: Incidencias

$$Riesgo - Rëlativo \Rightarrow \quad RR = \frac{I_1}{I_2} = \frac{a/n_1}{c/n_2};$$

$$Error - Estándar \Rightarrow \quad EE_{(\ln RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} + \frac{1}{n_1} + \frac{1}{n_2}};$$

$$IC95\% \Rightarrow \quad RR \times e^{[\pm 1{,}96 \cdot EE_{(\ln RR)}]}$$

$$\text{Re}ducción - \text{Re}lativa - del - Riesgo \Rightarrow \quad RAR = (1 - RR)x100$$

$$Riesgo - Atribuible \Rightarrow RD = I_1 - I_2 = \left(\frac{a}{n_1}\right) - \left(\frac{c}{n_2}\right);$$

$$Error - Estándar \Rightarrow \quad EE_{(RD)} = \sqrt{\frac{I_1 \cdot (1 - I_1)}{n_1} + \frac{I_2 \cdot (1 - I_2)}{n_2}};$$

$$IC95\% \Rightarrow \quad RD \pm 1{,}96 \times EE_{(RD)}$$

$$Número - Necesario - a - Tratar \Rightarrow \quad NNT = RD^{-1} = \frac{1}{RD}$$

Estudios de Casos y Controles

$$Odds_1\left(R/T+\right) = \frac{a/m_1}{b/m_2} = \frac{a}{b};$$

$$Odds_2\left(R/T-\right) = \frac{c/m_1}{d/m_2} = \frac{c}{d};$$

$$Odds - Ratio \Rightarrow \quad OR = \frac{Odds_1}{Odds_2} = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c};$$

$$Error - Estándar \Rightarrow EE_{(\ln OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}};$$

$$IC95\% \Rightarrow \quad OR \times e^{[\pm 1{,}96 \cdot EE_{(\ln OR)}]}$$

Por otra parte, en epidemiología también se utilizan las razones, que son un cociente cuyo numerador no está contenido en el denominador; carece igualmente de unidades y su rango oscila entre 0 (0/n) e infinito (n/0). En el ejemplo previo, la razón de masculinidad (hombre / mujer) en la muestra propuesta es de 98/53= 1,85 que se interpreta diciendo que en la muestra hay un promedio de 1,8 hombres por cada mujer —o bien 18 hombres por cada 10 mujeres; o bien 185 por cada 100—. Una forma característica de razón en epidemiología es la Razón de Riesgos o Riesgo Relativo (RR) que supone la razón entre dos valores de incidencia; para nuestro interés, la incidencia de personas que mejoran en el grupo de tratamiento entre la incidencia de personas que mejoran en el grupo control.

Un caso particular de razón es la denominada en inglés con el término 'Odds', con difícil traducción al castellano, que en ocasiones ha sido interpretada 'ventaja', 'nomio' o 'razón de complementarios', traducciones poco exitosas de modo que se ha venido optando por utilizar de forma generalizada la expresión Odds. La Odds consiste en dividir, en un grupo dado, el número de individuos que presentan una característica de interés por el número de individuos que carecen de ella. Por ejemplo, personas que han mejorado al recibir un tratamiento entre las personas que no han mejorado al recibir el mismo tratamiento. La odds, como tal, ha de ser entendida como una medida de frecuencia, similar a la proporción; pero cuando se establece el cociente entre dos odds obtendremos una Odds Ratio (OR) y en este caso estaremos ante una medida comparativa o de asociación. A diferencia de la probabilidad que es un cociente entre el número de casos favorables entre el número de casos posibles, la OR es una razón entre el número de casos favorables entre el número de casos desfavorables.

Tanto el RR como la OR oscilan entre 0 e infinito, siendo el valor nulo el 1, ya que en ese caso el numerador y el denominador son iguales. A medida que se aleja de 1 se incrementa la magnitud del efecto (asociación positiva), de modo que una ratio de 2 indica que hay dos veces más casos en el numerador que en el denominador, una ratio de 3,5 tres veces y media más, y así sucesivamente. Si decrece hacia 0 (asociación negativa) la interpretación es inversa, de modo que una ratio de 0,5 indica que hay dos veces más casos en el denominador que en el numerador (1/0,5= 2), una ratio de 0,02 indica que hay 5 veces más casos (1/0,2= 5).

Apliquemos ahora estas nociones a un ejemplo práctico (Tabla 5) y veamos cuáles son los índices del tamaño del efecto basados en riesgos y su interpretación (. Retomando nuestro ejemplo inicial, aunque se proponen datos figurados, tenemos 80 personas que participan en un ensayo clínico que han sido asignadas de forma aleatoria a dos grupos: 40 reciben un tratamiento experimental y 40 reciben un placebo conformando el grupo control. De las 40 personas que han recibido tratamiento, en 32 casos se han evaluado como éxito al finalizar el periodo de estudio, de modo que en este grupo se ha observado una incidencia del 80% de mejora terapéutica (I_1= 32/40 = 0,8; 0,8×100= 80%). Ahora bien, en el grupo control también se han observado 10 casos que han mejorado su respuesta de salud, de modo que la incidencia de mejora terapéutica en este grupo de un 25% (I_2= 10/40= 0,25). Primera observación a tener en cuenta: en todo estudio y en la práctica cotidiana se presentan casos con recuperaciones espontáneas, o mejor dicho, que no se pueden atribuir directamente al efecto de un tratamiento controlado. De este modo, el éxito del tratamiento experimental no es del 80%, ya que en este grupo también cabe la posibilidad de que se hayan producido mejorías espontáneas.

Diferencia de Riesgos (DR). También llamada Riesgo Atribuible (RA) o Reducción Absoluta del Riesgo (RAR). Expresa qué incidencia de éxito es debida a la intervención una

TABLA 5

Indicadores del tamaño del efecto de la familia de riesgos (Tablas 2×2).

Respuesta / Grupo	Éxito	Fracaso	Odds / Incidencia	DR o RAR	NNT	RR	ILR	OR
Tto (n_1=40)	32	8	O_1 = 4,00 I_1 = 0,80	0,55	1,8 ≈ 2	3,20	220%	12,00
Ctrol (n_2=40)	10	30	O_2 = 0,33 I_2 = 0,25	IC95% 0,36 a 0,74	IC95% 1,3 a 2,8	IC95% 0,21 a 6,19	----	IC95% 8,7 a 15,4

DR o **RAR**: Diferencia de Riesgos o Reducción Absoluta del Riesgo;
NNT: Número Necesario de pacientes a Tratar;
RR: Riesgo Relativo; **IRR**: Incremento del Logro Relativo; **OR**: Odds Ratio;

vez se controla el efecto de los casos que mejoran de forma espontánea y, por tanto, nos ofrece una medida de la frecuencia de mejoría que podemos 'atribuir' al tratamiento. Para su cálculo se resta a la incidencia de éxito terapéutico en el grupo de tratamiento la incidencia de éxito en el grupo control, de modo que en nuestro caso sería: DR= 0,80–0,25= 0,55. Es decir, la frecuencia de éxito terapéutico atribuible al tratamiento probado es del 55%; o dicho de otro modo, de cada 100 personas que fueran tratadas con este tratamiento mejorarían 55. La DR o RAR, varía de –1 a +1 (o de –100% al +100%, dependiendo de cómo queramos expresarla), siendo el 0 expresión de un efecto nulo (las incidencias de éxito se igualan en ambos grupos), los valores negativos expresarían que se han observado más casos con mejoría en el grupo control que en el experimental, y los valores positivos inclinarían la decisión a favor del grupo de tratamiento. Obtener una DR negativa, lo cual iría en contra de nuestras hipótesis, no significa que el resultado no sea relevante; simplemente, dependiendo de la magnitud del efecto, deberemos o bien revisar la ejecución de nuestro estudio, no sea que se hayan producido algunos acontecimientos que hallan hecho variar los resultados, o bien replantearnos nuestra hipótesis, lo cual nos conduciría a tomar una decisión clínica: el tratamiento propuesto no es mejor que el placebo, por lo tanto, mejor no utilizarlo ya que no añada nada favorable y si pudiera introducir efectos inconvenientes.

Por otro lado, estos índices también están sometidos a cierto grado de error de medida y, por tanto, son susceptibles de ofrecer un nivel de incertidumbre asociado a una significación estadística. No se apuren, que no vamos a calcularlo aquí, aunque cualquier programa estadístico nos lo ofrecerá sin problemas. Pero sí hemos calculado el intervalo de confianza del 95% de la DR (Tabla 5), que en nuestro caso se sitúa entre los valores 0,36 a 0,74. Observamos que el valor 0 no se halla comprendido en dicho intervalo y, por tanto, en ningún punto de dicho recorrido se hace el efecto nulo. Es decir, a falta de una prueba de significación estadística el IC95% nos ofrece información que nos permite decidir que la diferencia encontrada no se debe al azar sino que es producto del tratamiento, y que en el 95% de las ocasiones que realicemos este estudio el éxito terapéutico que lograremos alcanzar se situará entre un 36% y un 74%.

Número Necesario de personas a Tratar (NNT). Introducido por Laupacis, Sackett y Robers (1988), es un índice del TE muy utilizado recientemente. El NNT es un valor o indicador específico para cada tratamiento y des-

cribe la diferencia entre un tratamiento activo y un control (placebo) en lo que se refiere al logro de un resultado clínico concreto. Matemáticamente es el reciproco de la DR (NNT = 1/DR), que en nuestro caso es de 1,8 (1/0,55). Un NNT de 1 significa que en cada paciente al que se le da tratamiento se produce un resultado favorable, a la vez que ningún paciente del grupo de comparación (tratamiento control o placebo) tiene este resultado; es decir, todos los casos tratados mejoran, mientras que ninguno de los no tratados logra mejorar. Si bien no se han establecido límites para que un NNT sea reconocido como clínicamente efectivo, se acepta que el NNT es mejor cuanto más bajo sea su valor. Por lo general, el valor NNT se redondea al valor entero más inmediato, en nuestro caso a 2, y esto significa que debemos tratar a dos personas para que una se beneficie adicionalmente comparado con el grupo control. El NNT indica el esfuerzo 'adicional' que se requiere para conseguir un determinado efecto terapéutico, de modo que a medida que aumenta su valor mayor será el esfuerzo a realizar y, por tanto, comprometerá en mayor medida la decisión clínica. Al igual que el resto de índices, el NNT es una estimación sometida a incertidumbre y, por tanto, es conveniente presentarlo con su IC95% asociado (Pita y López–Ullibarri, 1998).

El concepto de NNT también se puede utilizar para valorar riesgos, por ejemplo los efectos adversos que pudiera causar un determinado tratamiento. Es decir, si una determinada intervención produce daños, y podemos calcular la incidencia de estos daños en ambos grupos, entonces podríamos estimar el NND o Número Necesario para Dañar (NNH–Number Necessary to Harm) con el mismo algoritmo que el NNT, es decir, el reciproco de la DR de las incidencias de efectos dañinos en ambos grupos y, en este caso, con un redondeo hacia el valor inmediatamente inferior. En este caso, a mayor NND mejor será la intervención ya que haría falta tratar a mucha gente antes de que apareciera un efecto adverso.

Riesgo Relativo (RR). Como ya ha sido descrito más arriba, se trata de una razón entre dos riesgos, entre dos incidencias acumuladas. Expresa por cuánto se multiplica la probabilidad del efecto, del éxito terapéutico, en personas que reciben el tratamiento en comparación con las que no lo reciben. Para conocer la precisión de las estimaciones aquí también es necesario calcular el IC95%. Como ya se ha planteado, éste nos informa sobre la significación estadística de la estimación del RR, de modo que si el IC incluye el valor de la hipótesis nula, es decir la incidencia de éxito entre los que reciben tratamiento y los que no es la misma y, por tanto, correspondería con el valor unidad (si $I_1=I_2$; entonces $I_1/I_2= 1$), entonces, decimos, no se podrá concluir que el efecto observado es real ya que pudiera ser debido al error aleatorio. En el ejemplo propuesto (Tabla 5), el valor de RR es de 3,20 con IC95% de 0,21 a 6,19. Es decir, la probabilidad de lograr un éxito terapéutico se multiplica por 3,20 entre los que han recibido el tratamiento en comparación con los que han tenido un placebo. También puede plantearse como que el éxito terapéutico es 3,20 veces más frecuente entre las personas sometidas a tratamiento que entre los controles. Como el IC no comprende el valor nulo (la unidad), estos resultados son estadísticamente significativos, es decir, es poco probable que el efecto sea debido al azar y, por tanto, podemos inclinarnos por considerar el tratamiento a favorable.

Reducción del Riesgo Relativo (RRR). Este índice está asociado al RR y pretende describir el TE en términos porcentuales, lo que resulta más cómodo e intuitivo para el clínico no formado en metodología. Para su cálculo basta restar a la unidad el valor RR y multiplicar dicha sustracción por cien [(1–RR) × 100]. Tratándose de evaluación del éxito terapéutico, el resultado que esperamos obtener no se trata tanto de un riesgo (probabilidad de que se produzca un determinado efecto adverso) sino de un logro, por lo tanto, hay que acomodar la fórmula del modo siguiente: [(RR–1) × 100], y

en este caso hablaríamos de *Incremento del Logro Relativo (ILR)*, si se nos permite la expresión. En nuestro ejemplo, para un RR= 3,20, el resultado del ILR sería de 220, es decir, la probabilidad de tener un éxito terapéutico entre los que reciben el tratamiento frente los controles sería de un 220% mayor.

Odds Ratio (OR). Este, junto al RR, son los más clásicos y utilizados de los índices de riesgo. La OR es un índice más comúnmente utilizado en los estudios de Casos y Controles, en los cuales no es posible calcular incidencias y, por tanto, el DR y RR. El diseño de este tipo de estudios parte de la clasificación de los grupos en función de la presencia o no del desenlace y, por tanto, no son adecuados para valorar resultados terapéuticos. No obstante, si es posible calcular los OR en estudios de intervención, y ofrecernos información relevante para la interpretación de los resultados. Como se explicaba al inicio de este apartado, la OR es una razón de odds, y ésta consiste en un cociente entre el número de individuos que presentan una característica de interés y el número de individuos que carecen de ella. Siguiendo nuestro ejemplo (Tabla 5), la odds de éxito en el grupo de tratamiento es de 4 (32/8) y la odds de éxito en el grupo control es de 0,33 (10/30), y dado que la Odds Ratio es la razón de odds, entonces tenemos que OR= 4/0,33= 12. Este resultado se interpreta de la forma siguiente: por cada caso control que ha presentado un éxito terapéutico hay 12 casos tratados experimentalmente que han mejorado. Es decir, el éxito terapéutico ha sido 12 veces más frecuente entre las personas que han recibido tratamiento que entre los que han recibido placebo. Esto nos conduciría a afirmar que probablemente el tratamiento es mejor que el placebo, y de hecho podemos afirmarlo ya que el intervalo de confianza del 95% (8,7 a 15,4) no contiene el valor nulo (OR=1) y, por tanto, el efecto encontrado es estadísticamente significativo.

Como podemos apreciar, el valor de la OR tiende a alejarse del RR, es decir, tiende a sobreestimarlo; de hecho, la OR es el producto de los RR posibles. Es decir, el RR es de 3,20 (32/10) si se comparan las tasas de éxito, y de 3,75 si se comparan las tasas de fracaso (30/8); y como puede apreciarse 3,20×3,75= 12. Por tanto, debido a esta sobreestimación, la OR siempre nos dará valores más protectores (OR<1) o de mayor riesgo (OR>1) que sus correspondientes RR. En consecuencia, la OR puede ser muy engañosa como índice del tamaño del efecto en la valoración de la relevancia clínica (Irala et al, 2004; Lede y Copertari, 2008).

Interpretación de la magnitud del efecto a partir de los Intervalos de Confianza

En la entrega previa de esta serie (Iraurgi, 2009), abordamos en uno de sus apartados el papel de los intervalos de confianza como sustitutos a las pruebas de significación estadística, ya que esta herramienta aporta información sobre la magnitud y la precisión del efecto (Gardner y Altman, 1986; Candia y Caiozzi, 2005).

El Intervalo de Confianza (IC) construido a partir de una muestra, es un rango de valores mínimo y máximo entre los cuales esperamos que se encuentre el verdadero valor del parámetro que tratamos de estimar. Un intervalo de confianza del 95% quiere decir que si se toman 100 muestras de un mismo tamaño y se utiliza cada muestra para construir un IC del 95%, se podría esperar que en promedio 95 de los intervalos incluirían el verdadero efecto de la terapia y cinco no lo hicieran. Una de las características del IC es que existe una relación entre éste y la prueba de hipótesis: cuando el IC del 95% no contiene el valor '0' (en el caso de diferencias de medias) o el valor '1' (en el caso de la razón de riesgos) se presenta una diferencia estadísticamente significativa ($p<0,05$), mientras que si el IC contiene el '0', o el '1' según el caso, entonces no existirían efectos significativos estadísticamente ($p>0,05$). Por ello, el IC a demás de utilizarse como estimador de la magnitud y precisión, también es válido como medida de significación estadística.

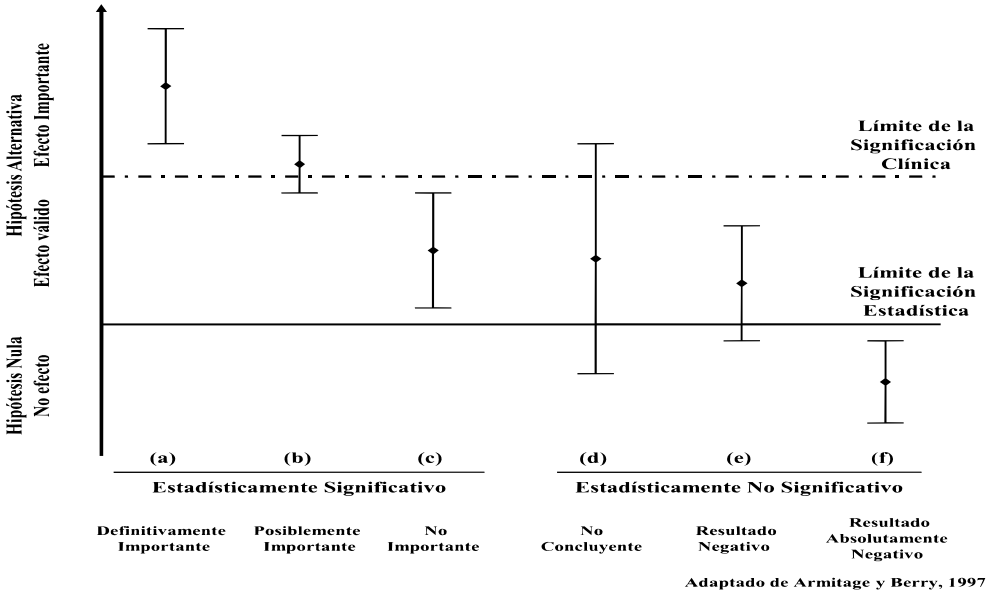
No obstante, a diferencia de los valores 'p', los intervalos de confianza no reducen los resultados a un simple 'blanco o negro', 'estadísticamente significativo o no significativo', sino que nos ofrecen un límite inferior y otro superior entre los cuales se sitúa el verdadero efecto en la población, es decir, nos ofrece una aproximación a la estimación del efecto. En el ámbito clínico, y en el científico en general, es preciso conocer si los resultados obtenidos en nuestra investigación pueden ser extrapolables a la población con un riesgo mínimo de equivoco —con una probabilidad inferior a un 5%—, pero también nos interesa sobre manera cuál es la magnitud del efecto logrado o la importancia de un resultado.

En la figura 2, adaptada de una propuesta de Armitage y Berry (1997), se presentan diferentes casos de efectos logrados y sus correspondientes intervalos de confianza. Se han clasificado en función de su posible interpretación en términos de significación estadística e importancia clínica. La figura es suficientemente intuitiva como para entenderla sin entrar en una descripción detallada. Pero quiero hacer notar el caso del ejemplo 'd'. En el observamos un efecto positivo (se halla por encima del límite de efecto nulo, barra horizontal continua) y ambos límites del intervalo de confianza sobrepasan ambas demarcaciones de decisión. Supongamos que el efecto es un RR de 2,20, que su IC95% oscila entre los valores 0,90 a

Figura 2.- Intervalos de Confianza e interpretación de la significación

Los Intervalos de Confianza (IC) muestran seis interpretaciones posibles en términos de significación estadística e importancia clínica:

(a) El efecto es estadísticamente significativo y lo suficientemente grande para tener con seguridad importancia clínica;
(b) La diferencia es significativa, pero no está claro si es lo suficientemente grande como para ser clínicamente importante (el límite inferior queda dentro del área de no relevancia clínica);
(c) La diferencia es significativa, pero demasiado pequeña para ser importante;
(d) La diferencia no es estadísticamente significativa pero puede ser suficientemente grande para ser importante (el límite superior del IC alcanza valores de significación clínica);
(e) La diferencia no es significativa ni tampoco lo bastante grande para ser clínicamente importante;
(f) La diferencia no es significativa ni tampoco lo bastante grande para alcanzar un efecto válido



Adaptado de Armitage y Berry, 1997

4,50, y que el valor de 4,0 es considerado como una puntuación clínicamente relevante. El límite del intervalo inferior se sitúa en un valor inferior al de significación estadística (0,90<1), lo cual lleva a la decisión de que el efecto logrado (en un mayor número de 5 veces de cada 100) pudiera ser debido al azar, es decir, no alcanza significación estadística. Ahora bien, el límite superior rebasa la zona de significación clínica, es decir, si el estudio se realizase 100 veces, en más de 5 veces de cada cien se presentarían tamaños del efecto superiores a 4, e incluso algunos llegarían has valores de 4,5. Probablemente este estudio no haya alcanzado la significación estadística por carecer de muestra suficiente —recordemos que el tamaño de la muestra repercute en la precisión de medida y esto contribuye a que la significación estadística pueda llegara apreciarse (Iraurgi, 2009)—, pero animaría a ser replicado con una planificación y diseño más depurado. Decidir que este estudio no es válido o no aporta información porque no resulta estadísticamente significativo nos llevaría a un gran error, ya que podría tener gran relevancia clínica pero que no ha podido ser apreciada por motivos, quizá, de diseño, de algún sesgo inoportuno, etc. Según el planteamiento de Kirk (1996) en ocasiones, un resultado que en un contraste de hipótesis ha sido 'no significativo' puede tener, sin embargo, una significación práctica. Una estimación puntual o un intervalo de confianza (la consideración de la magnitud del efecto, en definitiva) pueden usarse para decidir si los resultados son realmente pobres o, por el contrario, útiles e importantes.

Para concluir

Sin duda alguna, el tema de los indicadores del tamaño del efecto no se agota con lo expuesto en esta revisión. Existen gran número de trabajos y conexiones vía web (consúltense las referencias que aparecen con conexiones http) que permitirán al lector interesado ampliar mucho más sobre esta tema.

Nuestro propósito ha sido ofrecer una serie de índices del tamaño del efecto y abogar por la utilización de éstos y de los intervalos de confianza como herramientas útiles en la toma de decisiones cuando realizamos evaluación terapéutica o investigación en general. Hemos visto como el tamaño del efecto permite una apreciación más aproximada de la verdadera magnitud de los efectos de una intervención, y permite, asimismo, una interpretación más adecuada de los resultados. Por otro lado, y dado el auge que en las últimas décadas está tomando la medicina basada en la evidencia, cada vez se hace más necesario el cálculo y comunicación del TE obtenido en nuestros estudios, lo cual permitirá la integración de resultados mediante técnicas meta–analíticas. De ahí las recomendaciones de los expertos (Wilkinson y APATF, 1999) en que a la hora de informa de nuestros resultados aportemos los datos básicos: número de sujetos, medias, desviaciones estándar, correlaciones, y otros índices del tamaño del efecto.

Cerramos el artículo con las mismas preguntas que nos hacíamos al inicio. Cuando nos interesamos por la evaluación terapéutica, y la investigación en general (Kirk, 2001), tres son las preguntas básicas al examinar: 1) ¿se ha observado un resultado real o puede haber sido producto de casualidad, es decir, debe ser atribuido al azar?; 2) si el resultado es real, es decir, la estadística nos permite concluir que el resultado obtenido sólo se produciría por azar en muy pocas ocasiones (por debajo de 5 de cada cien veces), en ese caso, ¿cómo es de grande ese resultado?, o dicho de otro modo, ¿cuál es el tamaño o magnitud de su efecto?; y 3) ¿el resultado es lo suficientemente grande como para ser importante y útil?, es decir, ¿tiene significación clínica o práctica?. La aproximación a los conceptos para dar respuesta a la primera cuestión fue objeto de desarrollo en la primera entrega de esta serie metodológica (Iraurgi, 2009); la segunda cuestión ha sido abordada en la presente revisión, y dejaremos para el próximo número de la revista NORTE la dilucidación de la tercera. Allí les espero.

Contacto
Ioseba Iraurgi
Deusto-Salud, I+D+i en Psicología Clínica y de la Salud.

Universidad de Deusto.
Avda. Las Universidades. Bilbao
IRAURGI@telefonica.net

BIBLIOGRAFÍA

- Altman DG, Schultz KF, Moher D et al. (2001). The revised CONSORT statement for reporting randomized trials. *Ann Intern Med* 134:663–694.
- Argimon JM y Jiménez-Villa J (2004). Métodos de investigación clínica y epidemiológica. Tercera edición. Barcelona: Elsevier. Accesible en: http://books.google.es/books?id=_BLemLvp9XAC&pg=PA257&lpg=PA257&dq=clinicamente+significativo+relevante&source=web&ots=k6vBs5TQCL&sig=u32SczY0z6CCIRa9zS2CMQJxbn0&hl=es&sa=X&oi=book_result&resnum=9&ct=result#PPP1,M1
- Armitage P, Berry G. (1997). Estadística para la Investigación Biomédica. 3ª Edic. Harcourt Brace.
- Candia R, Caiozzi G. (2005). Intervalos de confianza. *Rev Méd Chile*, 133, 1111-1115. http://www.scielo.cl/pdf/rmc/v133n9/art17.pdf
- Carson C. (2004). The effective use of effect size indices in institutional research. Keene State College. http://www.keene.edu/ir/effect_size.pdf
- Clark ML. (2004). Los valores de P y los intervalos de confianza. *Rev Panam Salud Publica; 15(5): 293-296.*
- Coe R. (2002). It´s the effect size, Stupid. What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002. http://www.leeds.ac.uk/educol/documents/00002182.htm
- Cohen J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.
- Cohen J. (1969). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Laurence Erlbaum (2nd Edition, 1988).
- Cohen J. (1990). Things I have learned (so far). *American Psychologist*; 45(12): 1304-1312. Traducción: Cohen J. (1992). Cosas que he aprendido (hasta ahora). *Anales de Psicología*, 8, 1-2, 3-17. Accesible en: http://www.um.es/analesps/v08/02-08.pdf
- Cohen J. (1994). The earth is round (p<.05). *American Psychologist*; 49(12): 997-1003.
- Cooper HM, Hedges LV. (1994). *The Handbook of Research Synthesis*. New York Sage.
- Frías MD, Pascual J, García JF. (2002). La hipótesis nula y la significación práctica. *Metodología de las Ciencias del Comportamiento*; 181-185.
- Gardner MJ, Altman DG. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal*; 292: 746-750.
- Glass GV, McGaw B, Smith ML. (1981). *Meta-analysis in Social Research*. Beverly Hills: Sage Publications.
- González-Ramírez MT, Botella J. (2006). Comparación entre índices de tamaño del efecto para variables dicotomizadas en Meta-análisis. Psicológica, 27: 26-293. http://redalyc.uaemex.mx/src/inicio/ ArtPdfRed.jsp?iCve=16927207&iCveNum=7072
- Iraurgi I. (2000). Cuestiones metodológicas en la evaluación de programas terapéuticos. *Trastornos Adictivos*, 2, 2, 99-113. http://www.doyma.es/revistas/ctl_servlet?_f= 7016&articuloid=10017604
- Iraurgi I. (2009). Evaluación de resultados clínicos I.: Entre la significación estadística y la relevancia clínica. *NORTE de Salud Mental*, 33, 94-108. http://www.ome-aen.org/NORTE/33/NORTE_33_140_ 94-108.pdf
- Irala J, Martínez-González MA, Seguí-Gómez M. (2008). Epidemiología aplicada. Barcelona: Ariel.
- Kazdin AE, Bass D. (1989). Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, 57,138-147.
- Kirk RE. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Kraemer HC. (1992). Reporting the size of effects in research studies to facilitate assessment of practical or clinical significance. *Psychoneuroendocrinology* 17:527-536.
- Kraemer HC, Morgan GA, Leech N, Gliner JA, Vaske JJ, Harmon RJ. (2003). Measures of clinical significance. *J Am Acad Child Adolesc Psychiatry,* 42(12), 1524-1529.
- Laupacis A, Sackett DL, Roberts RS. (1988). An assessment of clinically useful measures of the consequesnces of treatment. N Engl J Med, 318, 1728-1733.
- Lede R, Copertari P. (2008). Indicadores básicos del efecto de las intervenciones médicas. Instituto Argentino de Medicina Basada en las Evidencias. http://www.iambe.org.ar/Indicadores.pdf
- Ledesma R, Macbeth G, Cortada N. (2008). Tamaño del efecto: Revisión teórica y aplicaciones con el sistema estadístico Vista. Revista Latinoamericana de Psicología, 40(3), 425-439. http://www.scielo.org.co/pdf/rlps/v40n3/v40n3a03.pdf
- Levy P. (1967). Substantive significance of significant differences between two groups. *Psychological Bulletin*, 67, 37-40.
- Lipsey MOW. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. London: Sage.
- Marín-Martínez F, Sánchez-Meca J. (1996). Estimadores del tamaño del efecto en meta-análisis; Un estudio Monte Carlo del sesgo y la eficiencia. [Effect size estimators in meta-analysis: A Monte Carlo study of bias and efficiency.] *Psicológica*, 17, 467-482. http://www.um.es/facpsi/ metaanalysis/pdf/7022.pdf
- Marín-Martínez F, Sánchez-Meca J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *Spanish Journal of Psychology*, 2, 32-38. http://www.um.es/ facpsi/metaanalysis/pdf/7040.pdf
- McGuigan FJ. (1993). *Experimental Psychology: Methods of Research* (6th Ed.). New York: Prentice-Hall.
- McQuay HJ, Moore RA. Using Numerical Results from Systematic Reviews in Clinical Practice. *Ann Intern Med* 1997; 126: 712-720. http://www.annals.org/cgi/content/full/126/9/712
- Pita Fernández S, López de Ullibarri Galparsoro I. (1998). Número necesario de pacientes a tratar para reducir un evento. *Cad Aten Primaria*; 96-98. http://www.fisterra.com/mbe/investiga/5nnt/5nnt.asp
- Pita-Fernández S, Pértega S. (2001). Significación estadística y relevancia clínica. Cad Aten Primaria, 8: 191-195. Accesible en: www.fisterra.com/ Accesible en:
- Primo J. (2003). Índices de eficacia de un tratamiento. NNT (II/II). *Enfermedad Inflamatoria Intestinal al día - Vol. 2 - Nº. 3 -62-66*. http://www.svpd.org/mbe/nnt.pdf
- Quezada C. (2007). Potencia estadística, sensibilidad y tamaño del efecto: ¿un nuevo canon para la investigación?. *Onomázein* 16(2): 159-170. http://onomazein.net/16/potencia.pdf
- Rosenthal R. (1991). *Meta-analytic Procedures for Social Research* (rev. ed.). Newbury Park, CA: Sage.
- Rosenthal R, Rubin DB. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166-169.
- Rosnow RL, Rosenthal R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- Rosnow RL, Rosenthal R. (1996). Computing contrasts, effect sizes, and counter-nulls on other people's published data: General procedures for research consumers. *Psychological Methods*, 1, 331-340.
- Sánchez-Meca J, Ato M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación. En J Arnau y H Carpintero (Coords.), *Tratado de Psicología General*, Vol. I (pp. 617-669). Madrid: Alhambra.
- Sánchez-Meca J, Marín-Martínez F. (2001). Meta-analysis of 2x2 tables: Estimating a common risk difference. *Educational and Psychological Measurement*, 61, 249-276. http://www.um.es/facpsi/ metaanalysis/pdf/7060.pdf
- Sánchez-Meca J, Marín-Martínez F. (2008). Confidence intervals for the overall effects size in random-effects meta-analysis. *Psychological Methods*, 13, 31-48. http://www.um.es/facpsi/ metaanalysis/pdf/5014.pdf
- Sánchez-Meca J, Marín-Martínez F, Chacón-Moscoso S. (2003). Effects size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, 8, 448-467. http://www.um.es/ facpsi/ metaanalysis/pdf/7078.pdf
- Wilkinson L, American Psychological Association Task Force on Statistical Inference (1999), Statistical methods in psychology journals. Guidelines and explanations. *Am Psychol*, 54, 594-604.